Method for automatically matching graphic elements and phonetic elements

The present invention relates generally to the automatic extraction of linguistic knowledges in a corpus of transcriptions of graphic chains into phonetic chains. It relates more particularly to the transcription of typographic elements such as characters in a predetermined language into phonetic elements.

At present, each word of a language constitutes a graphic chain that is transcribed phonetically into a chain of phonemes by a phonetician. For any new word to be added to a training corpus, the phonetician must intervene to transcribe the new word phonetically. Thus the training corpus furnishes only global grapheme/phoneme transcriptions. For example in the global transcription "ruelle"/[ryɛl], the corpus indicates that, globally, the graphic chain "ruelle" is translated into a phonetic chain. However, it is not made explicit that the typographic element "r" is retranscribed phonetically in some unitary way. The global transcription does not indicate also the syllables or graphemes constituting the graphic chain and the phonetic elements constituting the phonetic chain.

One or more phonetic chains associated with any graphic chain can be determined from the known elementary transcription of each typographic element by character by character analysis of the graphic chain. Error corrector systems find the phonetic transcriptions useful for recognizing lexical errors in entering text on a keyboard. There is therefore a need to extract more refined elementary transcriptions from a raw

transcription.

The invention aims to derive automatically from raw transcriptions of graphic chains, for example words and family names, into phonetic chains, transcriptions of graphic elements, for example characters, into phonetic elements constituting the phonetic chains, in order to segment any graphic chain into graphemes and any phonetic chain into phonemes automatically. The graphic element by graphic element, i.e. character by character, elementary transcriptions thereafter facilitate automatic global transcription of any additional graphic chain added to the corpus of graphic chains, in particular on the basis of a concatenation of phonetic elements matching on a one to one basis to the characters of the additional graphic chain.

Accordingly, a method of the invention matches graphic elements constituting given graphic chains automatically to phonetic elements constituting corresponding phonetic chains after initially entering global transcriptions of the graphic chains into the phonetic chains into a database accessible by the computer and after estimating and storing in the database first probabilities of elementary transcriptions of graphic elements into respective phonetic elements. The method is characterized by the following steps:

for each transcription of a given graphic chain with M graphic elements into a corresponding phonetic chain with N phonetic elements, determining second probabilities of MN second transcriptions of M graphic chains successively concatenating the M graphic elements into N phonetic chains successively concatenating the N

phonetic elements, each as a function of a respective first probability and of the highest of three respective second probabilities determined beforehand, and

establishing and storing a link between the last elements of the graphic and phonetic chains of each second transcription and the last elements of the graphic and phonetic chains of the transcription relating to the highest of the three respective second probabilities in order for links established in an M×N matrix relative to the second probabilities to constitute a single path between last and first pairs of graphic and phonetic elements of the matrix in order to segment the given graphic chain into graphemes corresponding to respective phonemes segmenting the corresponding phonetic chain and to store the matches between the graphemes and phonemes in the database, the number of graphic elements in a grapheme being identical to the number of phonetic elements in the corresponding phoneme, in order for any new graphic chain to be transcribed automatically into a phonetic chain segmented into phonemes by means of the stored matches.

According to other features of the invention, the respective first probability for the determination of a second probability relating to a second transcription of a graphic chain concatenating m graphic elements into a phonetic chain concatenating n phonetic elements, with $1 \leq m \leq M$ and $1 \leq n \leq N$, relates to the last elements in the graphic chain with m graphic elements and the phonetic chain with n phonetic elements. The three respective second probabilities determined beforehand for the second transcription of the graphic chain with m graphic elements into the phonetic chain with n phonetic elements preferably and respectively relate to a second

transcription of a graphic chain with m-1 graphic elements into the phonetic chain with n phonetic elements, a second transcription of the graphic chain with m graphic elements into a phonetic chain with n-1 phonetic elements and a second transcription of the graphic chain with m-1 graphic elements into the phonetic chain with n-1 phonetic elements.

For example, the invention transcribes phonetically from the corpus of global transcriptions such as "ruelle"|[ryɛl] the graphic elements "r", "u", "e", "lle" into the respective phonetic elements [r], [y], [ɛ], [l].

The invention may be regarded as similar to a process of syllabation which, by analysis, decomposes a global transcription into elementary transcriptions and locally matches grapheme/phoneme subtranscriptions. The division into initial graphemes and phonemes and the biunivocal matching of each graphic element to each phonetic element of the divided phonemes is called grapheme|phoneme alignment. In the above example, the invention produces the following alignment:

"r"	"u"	"e"	"lle"
[r]	[y]	[ɛ]	[l**].

The symbol * denotes a mute and meaningless phonetic element.

Other features and advantages of the present invention will become more clearly apparent from the reading of the following description of preferred embodiments of the invention, given by way of nonlimiting examples and with reference to the corresponding appended drawings, in which:

- FIG. 1 shows an algorithm of the main steps of the automatic matching method of the invention; and

- FIG. 2 shows an algorithm of the substeps of a step of the automatic matching method for determining individual first probabilities.

As shown in FIG. 1, the method of the invention for automatically matching graphic elements and phonetic elements comprises main steps E1 to E11. For example, those steps are for the most part implemented in the form of software in a terminal, such as a personal computer or a mobile in a cellular radio communication network, and linked in particular to software system for orthographic correction of lexical errors which is insertable into a word processing system or a linguistic practice system. The terminal contains or is able to access a database of the type used in artificial intelligence. The database stores a corpus C of initial global transcriptions.

Initially, in the step E1, the global transcriptions (CG|CP) are constituted by pairs each matching a graphic chain CG such as a word in a predetermined language or a family name to a phonetic chain CP. These transcriptions are determined and entered by an expert in phonetics on a form displayed by the computer. The corpus C matches *a priori* graphic chains GC each composed of one or more typographic elements (characters) hereinafter called graphic elements g_i of an alphabet $G = \{g_1, \ldots, g_I\}$ with I elements in the predetermined language, where $1 \leq i \leq M$, to respective phonetic chains CP each composed of one or more phonetic elements p_j of an alphabet $P = \{p_1, \ldots, p_J\}$ with J phonetic elements, where $1 \leq j \leq J$ and $I \neq J$. However, the segmentation of the chain CG into syllables or into graphemes each comprising one or more graphic elements and the segmentation of the chain CP into phonemes each

comprising one or more phonetic elements are ignored at this stage.

The alphabets G and P typically comprise around 30 elements. There are therefore a total of 30 × 30 = 900 possible pairs of graphic elements and phonetic elements. In practice, the corpus C contains at least 100 000 global transcriptions of typographic chains CG into phonetic chains CP, which protects the invention from coarse errors in the estimation of probabilities, as discussed below.

In the step E2, first probabilities of elementary transcription $P(g_i|p_j)$ such that a graphic element g_i matches the phonetic element p_j are firstly estimated and stored in the database with the corpus C of global transcriptions.

The estimated values of the first probabilities are as far as possible close to respective maximum probability values required for the method of the invention operating by iterations to converge quickly without retaining local maxima.

The concatenated nature of the global transcriptions of the chains leads to the hypothesis of a correlation between the rank r_g of the graphic elements in a graphic chain CG and the rank r_p of the phonetic elements in the corresponding phonetic chain CP. For example, in the global transcription (beau|bo), it is more probable that the graphic element b, given its position at the beginning of the chain CG, translates to a phonetic element [b] rather than a phonetic element [o] placed at the end of the corresponding chain CP. In this example, the correlation of the ranks moves the graphic elements [b] and [e] of the phonetic element [b] and the graphic elements [a] and [u] of the phonetic element [o]

closer together.

The algorithm for the initial estimation E2 of the first probabilities $P(g_i|p_j)$ comprises the following substeps E21 to E27.

In the substep E21, IJ contingency numbers K_{gipj} respectively associated with the elementary transcriptions $(g_i|p_j)$ of a graphic element of the alphabet G and a phonetic element of the alphabet P are set to zero. The contingency number K_{gipj} is equal at the end of the step E2 to the estimated number of times that the graphic element g_i is retranscribed into the phonetic element p_j in the various global transcriptions of typographic chains CG into phonetic chains CP included in the corpus C.

For each chain transcription (CG|CP), as indicated in the substep E22, the ranks of the graphic elements in the chain CG and the ranks of the phonetic elements in the chain CP are normalized as a function of the respective lengths l_g and l_p of the chains CG and CP, which may be different. In the substep E23, the rank r of a phonetic element in the chain CP is derived from the rank r_{gi} of a graphic element g_i in the chain CG with which the phonetic element of rank r will be associated, in accordance with the following relationship:

$$r = \text{integer portion } (r_{gi}.l_p/l_g).$$

The number K_{gipj} of contingencies associated with the elementary transcription of the graphic element g_i into the phonetic element p_j is then incremented by 1 only if the phonetic element p_j is situated at the derived rank r in the chain CP, as indicated in the substeps E24 and E25.

The substeps E22 to E25 are repeated for each global transcription (CG|CP) of the corpus C, as

indicated in the substep E26. When all the global transcriptions of the corpus have been processed, the next substep 26 estimates all the first probabilities $P(g_i|p_j)$ of elementary transcription between the graphic elements and the phonetic elements, in accordance with the following relationship for each graphic element g_i:

$$P(g_i \mid p_j) = K_{gipj} \;/\; \sum_{j=1}^{j=J} K_{gipi}$$

after calculating the sum term in the denominator for the graphic element g_i.

Referring again to FIG. 1, the matching process continues with steps E3 to E10 which segment each graphic chain CG read in the corpus of the database in order to match automatically and on a biunivocal basis each segment of the chain CG, called a grapheme, comprising one or more graphic elements, to a segment, called a phoneme, comprising one or more phonetic elements resulting from segmentation of the corresponding phonetic chain CP.

A graphic chain CG comprises M consecutive graphic elements g_1 to g_M and the phonetic chain CP corresponding to the chain CG comprises N consecutive phonetic elements p_1 to p_N. The integer N may be different from or equal to the integer M.

The probability $P(g_1,\ldots g_m,\ldots g_M|p_1,\ldots P_n,\ldots p_N)$ that the chain CG matches the chain CP, where $1 \leq m \leq M$ and $1 \leq n \leq N$, is determined as a function of the first elementary transcription probabilities $P(g_i|p_j)$ estimated and stored beforehand in the step E2 and from similarity between the chains CG and CP. The similarity is based on the Damerau-Levenshtein Metric (DLM) but using maximization instead of minimization. The probability P(CG|CP) is determined by dynamic programming using the

following iterative formula for any pair m,n such that $1 \leq n \leq N$ and $1 \leq m \leq M$:

$P(g_1g_2...g_m|p_1p_2...p_n)=P(g_m|p_n)\max[P(g_1g_2...g_{m-1}|p_1p_2...p_n)$, $P(g_1g_2...g_m|p_1p_2...p_{n-1})$, $P(g_1g_2...g_{m-1}|p_1p_2...p_{n-1})]$.

The concatenated nature of the global chain transcriptions and the grapheme/phoneme transcriptions means that Markov models may be applied efficaciously. For the given probability of transcription of a chain $g_1,g_2...g_m$ into a chain $p_1p_2...p_n$, the extension of the graphic, respectively phonetic, chain by a new graphic element g_{m+1}, respectively p_{n+1}, gives rise either to the same phonetic chain, respectively graphic chain, or to the addition of a new phonetic element, respectively graphic element. Expressed in terms of probability, $P(g_1g_2...g_{m+1}|p_1p_2...p_{n+1})$ depends only on the probabilities of three possible transcriptions:

$P(g_1g_2...g_m|p_1p_2...p_{n+1})$,

$P(g_1g_2...g_{m+1}|p_1p_2...p_n)$,

$P(g_1g_2...g_m|p_1p_2...p_n)$.

That dependency is expressed by the DLM metric equal to the highest of the above three possibilities.

After setting the indices m and n to zero for a global transcription (CG|CP) in the step E3 and incrementing the indices m and n by 1 in the steps E4 and E5, iterations in the steps E6 and E7 begin by determining the probabilities so that the M successive concatenations of the graphic elements g_1 to g_M of the chain CG match the first phonetic element p_1 of the chain CP, i.e.:

$$P(g_1,...g_m|p_1) = P(g_m|p_1) \max[P(g_1...g_{m-1}|p_1)]$$

where $1 \leq m \leq M$, and starting with the elementary probability $P(g_1|p_1)$. As shown by the step E8, the process then determines by iteration the probabilities of

the M concatenations of the graphic elements g_1 to g_M of the chain CG matching the first two phonetic elements p_1 and p_2 of the chain CP using the probabilities previously determined for the first graphic element p_1, i.e.:

$P(g_1,\ldots g_m|p_1, p_2) = P(g_m|p_2) \max[P(g_1,\ldots g_{m-1}|p_2), P(g_1,\ldots g_m|p_1), P(g_1,\ldots g_{m-1}|p_1)]$.

The process then continues by adding a phonetic element p_n to determine the M probabilities $P(g_1|p_1,\ldots p_n)$ to $P(g_1,\ldots,g_M|p_1,\ldots p_n)$ up to the M probabilities relating to the chain CP = $(p_1,\ldots p_N)$. By iteration of the steps E4 to E8, the computer progressively constructs and stores a matrix of second probabilities $P(g_1,\ldots g_m|p_1,\ldots p_n)$ with M columns for successive concatenations of the M graphic elements and N rows for successive concatenations of the N phonetic elements, operating row by row as in the above example, beginning with the probability $P(g_1|p_1)$ and ending with the probability $P(g_1,\ldots g_M|p_1,\ldots p_N)$.

Each iteration relating to the $(m.n)^{th}$ transcription $[(g_1,\ldots g_m)|(p_1,\ldots p_n)]$ establishes a link between the pair (g_m,p_n) and the pair with the highest of the three probabilities determined beforehand for the three pairs (g_{m-1},p_n), (g_m,p_{n-1}) and (g_{m-1},p_{n-1}). The link is stored in the computer. If the pair (g_m,p_n) is linked to the pair (g_{m-1},p_n), it is an elementary transcription from (g_{m-1},g_m) to g_m; if the pair (g_m,p_n) is linked to the pair (g_m,p_{n-1}), it is an elementary transcription from g_m to (p_{n-1},p_n); if the pair (g_m,p_n) is linked to the pair (g_{m-1},p_{n-1}), it is an elementary transcription from g_m to p_n.

Thus a link is stored in the computer for each determination of a probability $P(g_1,\ldots g_m)|(p_1,\ldots p_n)$. The links trace a single path that is also stored progressively in the computer and links the first pair

(g_1, p_1) to the last pair (g_M, p_N) in the matrix with M columns and N rows. The topology of the single path in the M×N matrix segments the graphic chains CG into graphemes and the phonetic chains CP into phonemes and aligns the graphic elements and the phonetic elements in biunivocal correspondence. If a segment of the path follows a portion of a row between two graphic elements, the concatenation of the graphic elements of that row portion corresponds to the phonetic element of the row completed by one or more mute and meaningless phonetic elements in order to form a grapheme and phoneme pair that has the same number of elements and is stored in the computer. If a segment of the path follows a column portion between two phonetic elements, the graphic element of the column plus one or more meaningless graphic elements corresponds to the concatenation of the phonetic elements of that column portion in order to form a grapheme and phoneme pair that has the same number of elements and is stored in the computer. A change of direction of the path in the matrix towards the horizontal, the vertical or the diagonal indicates segmentation of the chains CG and CP.

A simple example concerns seeking to segment the global transcription of the word CG = "beau" into the phonetic chain CP = [bo] on the assumption that the step E2 estimated the following first individual probabilities in the corpus C:

P(b|b)=0.9 ; P(e|b)=0.1 ; P(a|b)=0.1 ; P(u|b)=0.1

P(e|o)=0.2 ; P(a|o)=0.1 ; P(u|o)=0.2 ; P(b|o)=0.1.

For the transcription (beau|bo) from the corpus, the M=4 iterations of the steps E5, E6 and E7 for each of the N=2 rows of the 4×2 matrix produce the following table:

p_n / g_m	b = g_1	e = g_2	a = g_3	u = g_4
[b] = p_1	0,9	←0,09	←0,09	←0,0009
[o] = p_2	↑0,09	↖0,18	←0,018	←0,0036

The symbol ← indicates that the pair (g_m, p_n) is linked to the pair (g_{m-1}, p_n); the symbol ↑ indicates that the pair (g_m, p_n) is linked to the pair (g_m, p_{n-1}); and the symbol ↖ indicates that the pair (g_m, p_n) is linked to the pair (g_{m-1}, p_{n-1}). The symbol ↖ associated with the transcription (be|bo) indicates that the latter has been derived and is therefore linked to the preceding transcription (b|b). The symbol ↖ indicates a segmentation boundary between grapheme and phoneme pairs. The following alignment is derived from this table:

b   eau

b   o**.

The symbol * designates a mute and meaningless phonetic element.

To perfect the matches between graphemes and phonemes and the matches between graphic elements and phonetic elements, preferably in the manner indicated by the step E11, the first probabilities $P(g_1|p_1)$ to $P(g_I|p_J)$ of the transcriptions of each of the graphic elements respectively into the J phonetic elements (step E2) and in particular the contingency numbers K_{g1p1} to K_{gIpJ} (substep E25) are again estimated as a function in particular of the ranks of the phonetic elements placed in the given phonetic chains CG that were segmented into phonemes in the preceding step E10. Second probabilities $P(g_1,...g_m|p_1,...p_n)$ of M×N second transcriptions of each global transcription of a given graphic chain with M graphic elements (CG) into a corresponding phonetic chain

(CP) with N phonetic elements are determined by executing the steps E3 to E10 in order for links to be established in the next step E10 between pairs (g_m, p_n) of a new matrix with M columns and N rows and consequently for a corrected path to link the last pair (g_M, p_N) to the first pair (g_1, p_1) in the new M×N matrix of second probabilities.

Thanks to the processing capacity and high processing speed of the computer, other iterative loops of steps E2 to E11 may be executed in the computer until the matching process converges, i.e. until the path established becomes constant from one loop to the next.

After segmentation of all the graphic and phonetic chains of the corpus G into graphemes and phonemes, the database stores all matches between graphic and phonetic elements and all matches between graphemes and phonemes for the whole of the processed corpus C.

Any new graphic chain added to the corpus can then be transcribed automatically into a phonetic chain segmented into phonemes, in particular with the aid of the matches previously established and stored in accordance with the invention, which progressively enriches the corpus in the database and increases transcription accuracy.

As already stated, the phonetic transcriptions are useful to orthographic error correction software systems that recognize lexical errors when entering text on a terminal keyboard. Thus when the new graphic chain added to the corpus is being entered on a terminal keyboard, the phonetic chain segmented into phonemes by means of the stored matches is used for orthographic correction of the new graphic chain entered.

The method of the invention may equally well be used as a tool for automatically generating SMS short messages from a text written in ordinary language. This necessitates a training corpus C the transcriptions whereof are adapted to the automatic generation of SMS messages and respectively match graphic chains CG, such as words and phrases, to phonetic chains CP whose "phonemes" are phonetically readable by any person who is not an expert in phonetics. For example, the corpus establishes the following matches (in French) between graphic chains and phonetic chains:

j'ai : G
air : R
occupé : OQP
cas : K.

Thus a new graphic chain entered in a terminal is automatically transcribed by the method of the invention into a phonetic chain segmented into phonemes that can be read by any person who is not an expert in phonetics by means of stored matches to be included in an SMS message. In the foregoing example, the French phrase "j'ai l'air occupé" entered on the terminal is transcribed automatically into the following short message to be transmitted by the terminal: Gl'ROQP, the "phonetic chains" [G], [l'], [R] and [OQP] being phonetically readable by any user who is not an expert in phonetics. Alternatively, the phonetic chains [G], [l'], [R] and [OQP] may be treated as phonetic elements to constitute a phonetic chain [Gl'ROQP].

The steps of a preferred embodiment of the method of the invention are determined by instructions of a computer program incorporated into a computer such as a

terminal, a personal computer, a server or any other electronic data processing system. The program automatically matches graphic elements constituting given graphic chains to phonetic elements constituting corresponding phonetic chains, after initially entering global transcriptions of the graphic chains into the phonetic chains into a database accessible to the computer and estimating and storing in the database first probabilities of elementary transcriptions of graphic elements into respective phonetic elements. The program includes program instructions which execute the steps of the method of the invention when said program is loaded into and executed in the computer, the operation whereof is then controlled by executing the program.

Consequently, the invention applies equally to a computer program adapted to implement the invention, in particular a computer program on or in an information medium. This program may use any programming language and take the form of source code, object code or an intermediate code between source code and object code, such as a partially compiled form, or any other form that may be desirable for implementing the method of the invention.